NPTEL

NPTEL ONLINE CERTIFICATION COURSE

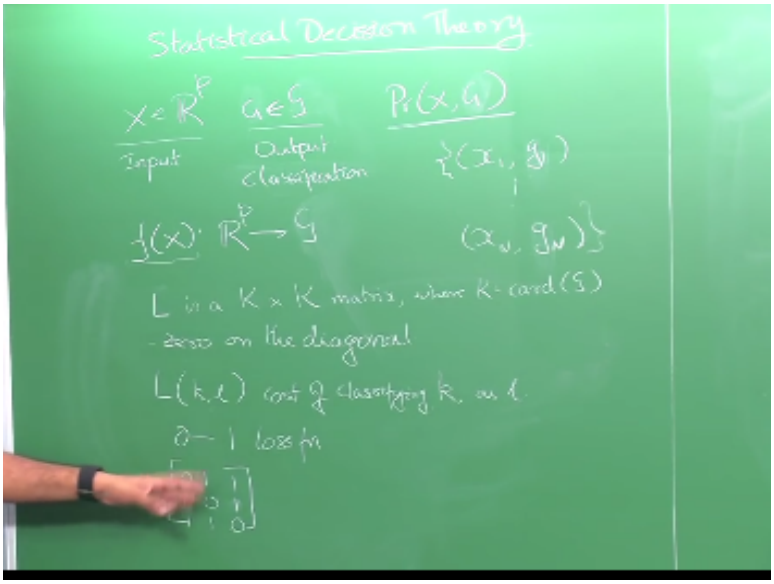
Introduction to Machine Learning

Lecture 10

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Statistical Decision Theory –
Classification

(Refer Slide Time: 00:16)



In this module we are going to look at the case where the output variable is drawn from a discrete space or in other words we are going to look at the classification problem as before the input is coming from a P dimensional space RP the output which I am denoting by G here I am going to assume is coming from some space script G which is a discrete value right it could be Bayer computer does not by computer so capital the script G could just consists of Bayer computer does not by computer or it should consist of like 5 different outcomes has the disease a mild form of the disease a severe form of the disease does not have the disease and soon so forth right.

So it could be a variety of outcomes but a small discrete set, so that space is denoted by script g well capital g is the random variable corresponding to the output right then like before we are going to have a joint distribution on the input on the output right and the training data is going to consist of pairs x_1 g_1 x_2 g_2 all the way up to x and Gn and the goal here is to learn a function f(x) that is going to take you from a P-dimensional input space R to the discrete space script g right.

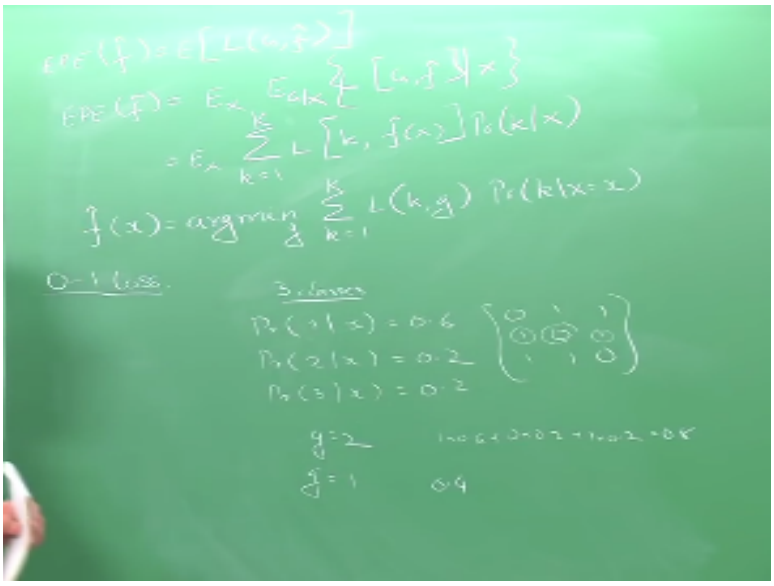
And so the thing that we have to look at now is what is an appropriate loss function in this case so what is an appropriate loss function in this case since we are talking about the discrete output right so I really cannot talk about squared error as a loss function even though in cases where the discrete values have been encoded as numeric outputs people do use squared error and we will see that later right so people do use squared error is an appropriate measure as long as your space g has been encoded numerically right.

So but in general so we are going to define the loss in as a k / k matrix where k is the cardinality of the discrete space script g that we are looking at, so suppose there are 5 classes then my last matrix is going to be a 5/ 5 matrix right so that the thing here is it is going to have 0 on 0 on the diagonal right and so the kl^{th} at entry in the last matrix essentially is the cost that you incur of classifying the output k as n so the true output is k but you output you say l right so that is essentially the cost of classifying k as l so that is denoted by the kl^{th} entry of the loss matrix right.

So frequently the most popular loss function that you use is known as the 0 - 1 loss function right so the 0-1 loss function essentially says that suppose I have three classes right, so my loss function would look like this right, so if I if I classified to the right class I get a penalty of zero but if I classify to the wrong class right I get a penalty of one regardless of which wrong class I classify too, so this entry says that okay the data point actually belongs to class one I have classified it as class two what is the penalty so 1 data point belongs to class 1 I classify it as class 3.

What is the penalty one and so on so forth so this is called the 0-1 loss function because all the entries in the loss matrix are either 0-1-1 right.

(Refer Slide Time: 04:26)



So what we are again going to look at is the expected prediction error of here the right and we can do the same thing that we did earlier so I can start conditioning it on x right so the expected prediction error and then the expectation of g / g given x which essentially becomes a right so the loss of g, f^ given that the input is x but if you think about it this is not a continuous distribution this is actually a discrete distribution because g can take only finitely many values, so in so writing it out as this expectation i can actually simplify that and write it as right.

So this is a loss that i will incur if k was the true class again my prediction was f^(x) times the probability that k is the true class given the input x right so this is essentially i am writing out the expectation here right, so the because it is a discrete distribution i am able to write it out in a compact form right and again i can do this minimization of this point twice like we talked about earlier, so point wise would mean that I make a specific assumption about what is the value of x right.

So I am going to look at right, so they are essentially following the same treatment that we did with the reduction case except that we are using a discrete output space since of a continuous output space right, so this essentially says that I am going to pick the data point right that gives me the I am going to prove sorry I am going to pick the prediction g that gives me the smallest expected error right, so what suppose I have the 0-1 loss function right assume the 0-1 loss so what does this mean I should essentially set my g to be that k right which has the highest probability why is that right.

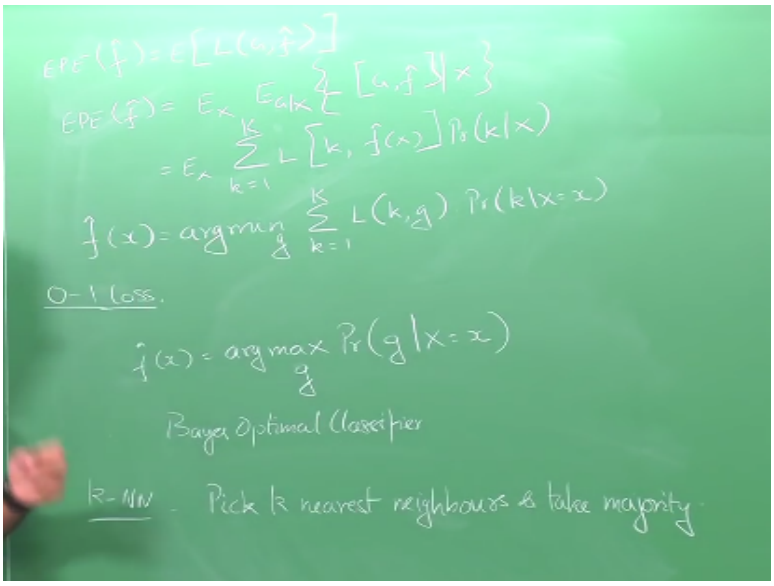
So if we think about it this probability term contributes to every element in the summation right so what I can do is among all these probability terms I can pick one term right and set it to 0 by my choice of g right so suppose I choose g to be 1 then my l(1,1) will become 0 right and but my l(2,1)(3,1) so on so forth will all be 1 so what will happen this there is a probability of 2 given X probability of 3 given x all of this will actually appear in this summation right.

So if I set my g to that value of k which has the highest probability right then that will yield the best possible solution here right, so if you are able to see that let us assume that there are 3 classes right I assume that there are 3 classes so and my true distribution is says that the probability of class 1 given the data point x let us say is 0.6 probability of + 2 given x let us say is 0.2 this is another 0.2 okay and of course my loss function is going to be such that and mean 00011111 right so this is my loss function right.

So if I guess that my class label is going to be 2 let us say so I said g = 2 so what is going to happen my if the class label is 1 right so I am going to look at the loss corresponding to 1, 2 which is this right, so I will get 1 times 0.6 then if the class label is 2 so I will be looking at this so I will get 0 times 0.2 + if the class label is 3 I look at this I will get 1 times 0.2, so I will get a score of 0.8 right.

So as you can see depending on which value I choose if I g = 2 then I will be zeroing out the second entering if I choose g = 1 I will be zeroing out the 1st entry right 4 by choosing g equal to 1 I will basically get a score of 0.4 right, so what I have to do in order to get the minimum here is to pick that g for which this probability is the highest right.

(Refer Slide Time: 11:31)



So I will set f^(x) okay so can you how people realize why the min here became the max here based on the argument that we just did right, so this is essentially saying that from your training data classify it to the most problem class right and if I knew this if I knew this probability right so what will I do I can set it to the most probable output so this is this kind of a classifier so what is the Bayer optimal classifier say I can look at the conditional distribution right given x look at the probability of g take the g that has the highest probability and assign it as the output so this is essentially what the Bayer optimal classifier would say all right.

But then you do not know g right, so what they have to do is you have estimate this probability so how would you estimate this probability do we know of any method for estimating this probability of course we do we know how to do nearest neighbor right, so what you would do in this case is that instead of taking the average over the neighbors like we did in the regression case so what you would do is you our estimate the probabilities in the neighborhood, so what you would do you will take a data point look at the k neighbors of the data point k nearest neighbors of the data point find out what their class labels are right.

So and then divide by up so for each label count the number of occurrences of that label in the k neighbors and divided by k right so this will give you the probability of the class label in the neighborhood but we really do not have to do this much work why because we are not interested in the actual probability we all we need is the one that has the maximum probability since the

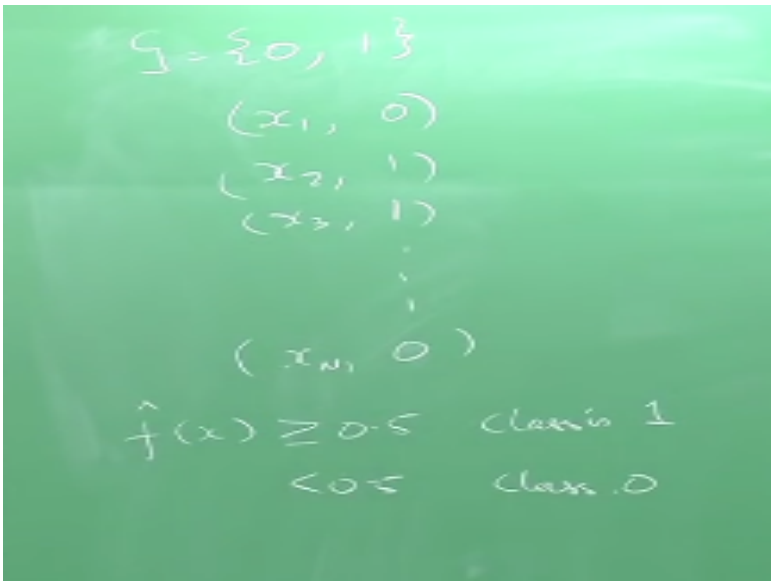
denominator is going to be k for all the probabilities we can ignore the denominator we can just look at the numerator.

So what we can do is we can count the occurrences of the class label in the neighborhood and whichever occurs more often we can assign that as the class label right, so think about it for a minute right, so what we are essentially doing when we take the majority is actually estimating this probability and taking the max probability right, so take the majority label in the neighborhood and use that as your prediction so this essentially gives you the k nearest neighbor classifier.

So what we saw earlier was a k nearest neighbor regress or so all the caveats that we talked about for the k nearest neighbor regress or appear when applied to the k nearest neighbor classifier as well so you have to be careful about using it in very high dimensions right and you really need large values of k and large values of n before you can get stable estimates but having said all that I should say that it turns out to be a really powerful classifier in practice and we will come back to that a little later as to why it is such a powerful classifier right.

And can we use linear regression or the linearity assumption here it turns out that you could use linear regression in almost directly for solving this problem so the way you do it is the following you take this data set x_1 g_1 x_2 g_2 and so on so forth and convert it into a data set suitable for doing regression, so how do I do that so I take that x_1 g_1 right.

(Refer Slide Time: 16:10)



Let us say that I have only two classes for simplicity sake let us say I have only two classes right so I have g1 and g2 right so let us say 1 and 2 so I will say I will say that 0 or 1 right so instead of having some arbitrary classes I am going to say it is 0 or 1 so what I am going to now do is my thing will leave that become something like this right so instead of having some arbitrary symbols g's g_1 g2 I am going to have 0, 1, 1, 0 and so on so forth now what I can do is I can solve this as a regression problem I can just solve this as a regression problem and whatever output I get I can read that as an estimate of the probability of g given x if you think about it right.

So probability of g = 1 given x so for the same value of x if there are multiple ones right suppose I the same value x occurs say 5 times in my training data 3 of the times it was 1 and 2 of the times it was 0 right so when I am trying to do a prediction I would expect to end up at the average of this prediction right, so just be like 3 / 5 and it also turns out to be the probability with which the output is 1given an x right, so if I do regression with this as my training data, so what I will be learning is the probability that g = 1 given x right roughly there are lot of caveats in this which we will look at when we do regression later obviously you cannot treat this directly as probabilities because the regression curve can become negative right.
So you cannot really treat it as probabilities but it just how it is useful intuition to have and so the output that you learn here so $\hat{f}(x)$ right in this case if it ≥ 0.5 then you say in the classes 1 if it is < than 0.5 you say the class is 0 let us say can use a linear regression to solve this as well so what we have done in this couple of modules is to look at a unifying formulation for classification and

regression problem so supervised learning problems and looked at a couple of different classifiers that arise out of making certain assumptions about classifies and regresses that arise out of making certain assumptions about the function that we are trying to learn right.

In the subsequent classes we will start looking at each of these in more detail starting off with linear regression we'll look at this different classifiers in greater detail thank you.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in